Ranking Candidate Genes for the Biomarker Development in a Cancer Diagnostics

Byung Soo Kim [1*], Inyoung Kim [2], Sunho Lee [3], Sun Young Rha [2]

[1] Department of Applied Statistics, Yon sei University, Seoul, Korea
[2] Cancer Metastasis Research Center, College of Medicine, Yonsei University, Seoul, Korea
[3] Department of Applied Mathematics, Sejong University, Seoul, Korea
[*] To whom correspondence should be addressed. E-mail:bskim@yonsei.ac.kr

Abstract

Recently, Pepe *et al.* (2003) employed the receiver operating characteristic (ROC) approach to rank candidate genes from a microarray experiment that can be used for the biomarker development with the ultimate purpose of the population screeni ng of a cancer. In the cancer microarray experiment based on n patients the researcher often wants to compare the tumor tissue with the normal tissue within the same individual using a common reference RNA. This design is referred to as a reference design or an indirect design. Ideally, this experiment produces n pairs of microarray data, where each pair consists of two sets of microarray data resulting from reference versus normal tissue and reference versus tumor tissue hybridizations. However, for certai n individuals either normal tissue or tumor tissue is not large enough for the experimenter to extract enough RNA for conducting the microarray experiment, hence there are missing values either in the normal or tumor tissue data. Practically, we have n_1 pairs of complete observations, n_2 "normal only" and n_3 "tumor only" data for the microarray experiment with n patients, where $n=n_1+n_2+n_3$. We refer to this data set as a mixed data set, as it contains a mix of fully observed and partially observed pair data. This mixed data set was actually observed in the microarray experiment based on human tissues, where human tissues were obtained during the surgical operations of cancer patients.

Pepe *et al.* (2003) provide the rationale of using ROC approach based on two independent samples for ranking candidate gene instead of using t or Mann -Whitney statistics. We first modify ROC approach of ranking genes to a paired data set and further extend it to a mixed data set by taking a weighted average of two ROC values obtai ned by the paired data set and two independent data sets.

Introduction

Microarray technology holds the promise of becoming a valuable tool in cancer research and clinical diagnostics with its potential to

[1] This work was supported by Ministry of Health & Welfare and Korea Science and Engineering Foundation

quantitatively measure the expression levels of thousands of genes simultaneously. Two technologies are widely used, the cDNA microarray and the oligonucleotide array, which differ with respect to the types of nucleic acid probes arrayed for the interrogation of labeled RNA specimens. Here we focus on the cDNA microarray, which uses a dual-label system in which two RNA specimens are separately reverse transcribed, labeled, mixed and hybridized together to each array.

One of the characterizing properties of cDNA microarray data is that it is subject to substantial variability, hence it is essential for the experimenter to carefully plan the experimental design driven by the study objective. For example, when using a cDNA array one must decide on a design for allocating specimens to labels and to array. The most commonly used design uses an aliquot of a reference RNA as one of the specimens for each array. This design is often referred to as a reference design or an indirect design. For most cancer microarray experiment, as Simon *et al.* (2002) indicates, there is not enough material available from one individual to create multiple arrays and thus the reference design would be the design of the choice. It is also a common practice in the cancer microarray experiment that a normal tissue is collected during the surgery from the same individual from which the tumor tissue was taken. The major reason of doing this is that it is not easy for the experimenter to collect normal tissues from healthy individuals. Also by observing a matched pair from a same individual one can reduce the inter-individual variability. It is often the case, however, that the experimenter can't extract enough RNA either from the tumor or the normal tissue to perform the microarray experiment due to poor quality of the tissue or other technical reasons. Therefore, collecting n cases does not necessarily end up with a matched pair sample[2] of size n. Instead it usually consists of a matched pair sample of size n_1 and two independent samples of size n_2 and n_3, respectively for 'reference versus normal only' and 'reference versus tumor only' hybridizations ($n_1+n_2+n_3=n$). Let X and Y denote the log fluorescent intensity ratios of reference versus normal and reference versus turmor hybridizations, respectively. Let U and V be independent copies of X and Y, respectively. Then we may observe three data types represented in Table 1.

Table 1. Three data types of the experiment. X and Y represent log intensity ratios for reference versus normal and reference versus tumor hybridizations, respectively. U and V are identically distributed with X and Y, respectively.

Hybridization		Number of cases
reference vs normal	reference vs tumor	
X	Y	n_1
U	missing	n_2
missing	V	n_3

Experiment and Data pre-processing

The fresh specimens of cancer and normal tissues obtained from colorectal cancer patients during surgery were snap-frozen in liquid nitrogen right after the resection and stored at -70°C until required.

[2] We use 'sample' to denote a random sample in statistics to distinguish it from a biological specimen.

We originally attempted to extract total RNAs from tumor and normal tissues from 87 patients. From each of 36 patients we had RNA specimens both for tumor and normal tissues. However, from 19 patients RNA specimens for normal tissues only were available. From another 32 patients RNA specimens for tumor only were obtainable. Thus, we have a matched pair sample of size 36 and two independent samples of sizes 19 and 32. In terms of notations in Table 1 n_1=36, n_2=19 and n_3=32. After total RNAs were extracted from fresh frozen tissues, the specimens were labeled and hybridized to cDNA microarrays based on the standard protocol established at Cancer Metastasis Research Center (CMRC), Yonsei University College of Medicine. We used $M=\log_2(\frac{R}{G})$ for the evaluation of relative intensity, where R and G in (R, G) represent the cy5 and cy3 fluorescent intensities, respectively.

We first define no missing proportion (NMP) of a gene as the proportion of valid observations out of the total number of arrays. For example, if a gene has valid observations for 32 arrays out of 40, its no missing proportion is 0.8. We preprocessed the data as follows;

(1) We normalized the log intensity ratio, $\log_2(\frac{R}{G})$, using within-print tip group, intensity dependent normalization following Yang *et al.* (2002).

(2) We used 0.8 for the cut point of NMP to delete genes containing missing values for more than 20% of the total number of observations. This filtering procedure yielded 13859 genes.

(3) We employed k-nearest neighbor (k=10) method for the imputation of missing values.

(4) We averaged values for multiple spots. The numbers of duplicated, triplicated, and quadruplicated spots were 982, 6, and 5, respectively.

(5) Finally, we have a data set represented by a 12850 x 123 matrix, where 12850 represents the number of genes and 123 stands for the number of microarrays.

We investigated various box plots (not shown) after the (location parameter) normalization and concluded that it was not necessary to have the scale normalization either between blocks within an array or between arrays.

Methods

Recently, Pepe *et al.* (2003) employed the receiver operating characteristic (ROC) approach to rank candidate genes that can be used for the biomarker development with the ultimate purpose of the population screening of a cancer. Identifying DE genes in a cancer can be accomplished using various statistical methods, for example, those described in Kim *et al.* (2004). If a gene is found to be differentially expressed in the cancer, then by developing a suitable biomarker, the corresponding protein product or an antibody to it can be detected in blood or urine, which forms the basis for a population screening. There are two points that differ from identifying DE genes between normal and tumor tissues in the development of biomarker used for the population screening. First, clinical bioassays for some gene products may be too difficult to develop for technical reasons. Thus, we need to have a sizable number of candidate genes with development priorities so that if one gene proves to be useless for biomarker development, we may still explore the next gene for the development.

The second point is that the bioassay, once it is developed, is applied to the whole population, and hence the false positive rate should be extremely low. Even a small false positive rate yields a large number of healthy people being subjected to diagnostic procedures that are unnecessary, costly and sometimes invasive. For ranking candidate genes we need statistical measures which discriminate between normal and tumor tissues and the measure of choice should focus on separation of these two distributions. Pepe *et al.* (2003) provide the rationale of using ROC approach for this purpose of ranking candidate gene instead of using t or Mann-Whitney statistics. Let X and Y denote the log fluorescent intensity ratios of reference versus normal and reference versus turmor hybridizations, respectively. Let U and V be independent copies of X and Y, respectively. Then we may observe three data types represented by $\{(X_j, Y_j)\}_{j=1}^{n_1}$, $\{U_k\}_{k=1}^{n_2}$, and $\{V_l\}_{l=1}^{n_3}$ as in Table 1. We assume that $\{U_k\}_{k=1}^{n_2}$ and $\{V_l\}_{l=1}^{n_3}$ are independent random samples. Let U_1=U and let V_1=V. The ROC curve is a plot of true positive versus false positive probabilities associated with varying thresholds c for U and V. Just for the simplicity we present ROC curve for the up-regulated genes. However, the adaptation to the down-regulated gene is straightforward. For a given threshold c the false positive probability is given by Pr[U≥c]≡t, and the true positive probability is Pr[V≥c]≡ROC(t). Therefore, the ROC curve consists of {(t, ROC(t)); 0≤t≤1 }. Pepe *et al.* (2003) used empirical estimates of $ROC(t_0)$ together with $pAUC \equiv \int_0^{t_0} ROC(t)\, dt$ for ranking genes of differential expression between normal and tumor tissues for a suitably chosen t_0 (≡ $Pr[U \geq c_0]$) value.

The ROC approach is now extended to the mixed data set of Table 1. We introduce some notations here. Let $ROC_{pair}(c_0)$, $ROC_{ind}(c_0)$ and $ROC_{mix}(c_0)$ denote ROC values using the matched pair sample, two independent samples, and the mixed data set of Table 1, respectively, for a given threshold c_0. Let D=Y-X and let D_0 denote the hypothetical version of D under the null hypothesis of no differential expression. The distribution of D with a mean δ_a and the variance σ_a^2 is denoted by $D \sim (\delta_a, \sigma_a^2)$. The distribution of D_0 is represented by $D_0 \sim (\delta_0, \sigma_0^2)$. We augment the D notation by adding a superscript (g) to represent the g-th gene. Hence $D^{(g)}$ denote D for the g-th gene. We omit this superscript when the argument is based on each gene. Let $|\overline{D}|_{(1)} \leq |\overline{D}|_{(2)} \ldots \leq |\overline{D}|_{(p)}$ denote the order statistics of $\{|\overline{D}^{(g)}|\}_{g=1}^{p}$, where p is the number of genes spotted in a cDNA microarray and $\overline{D}$ denotes the sample mean of D values based on n_1 observations.

The extension consists of estimating $ROC_{pair}(c_0)$ and $ROC_{ind}(c_0)$ and then averaging these two values to yield $ROC_{mix}(c_0)$. The essence of the extension is to estimate the baseline distribution of D under the null hypothesis of no differential

expression, which $D_0 \sim (\delta_0, \sigma_0^2)$ denotes. We assume, for simplicity, that D_0 has the same distribution with D except for the mean and variance. We expect that $\delta_0 \leq \delta_a$ and we don't necessarily assume that $\delta_0=0$. We further assume that $\sigma_0^2 \leq \sigma_a^2$. There are several ways of estimating the distribution of D_0 using the matched pair sample data. We choose a set of genes, denoted by $N_\varepsilon = \{g; |\overline{D}^{(g)}| < \varepsilon\}$ for a small $\varepsilon>0$. The suitable choice of ε can be determined from the plot of $\{(m, Var(|\overline{D}|_{(m)})\}_{m=1}^{p}$. Based on the genes in N_ε we can estimate δ_0 and σ_0^2, and proceed the calculation of $ROC_{pair}(c_0)$. Calculating $ROC_{ind}(c_0)$ is straightforward and hence we omit the details. Once $ROC_{pair}(c_0)$ and $ROC_{ind}(c_0)$ are determined we may average these two ROC values to get $ROC_{mix}(c_0)$. The initial idea was using the weighted mean of these two ROC's where the weights were proportional to the inverse of their variances. For calculating the variance of $ROC_{ind}(c_0)$ we attempted using Result 5.1 of Pepe (2003). We employed bootstrap method for calculating the variance of $ROC_{pair}(c_0)$. However, for some of genes which well separated distributions of normal and tumor tissues, we observed that zero variances occurred when we performed bootstrap procedure for calculating the variance of $ROC_{pair}(c_0)$. We also noted that Result 5.1 of Pepe (2003) didn't work out for computing variance of $ROC_{ind}(c_0)$, because for some genes which well separated two distributions Equation (5.2) of Pepe (2003) involved division by zero. Therefore, we used following weighted average to get $ROC_{mix}(c_0)$;

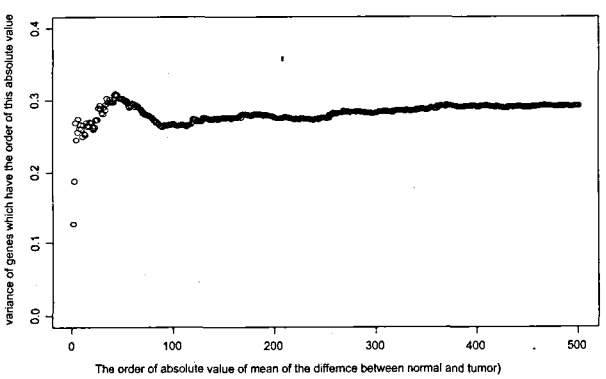
$$ROC_{mix}(c_0) = \frac{n_1 ROC_{pair}(c_0) + n_H ROC_{ind}(c_0)}{n_1 + n_H}, \quad (3)$$

where n_H is the harmonic mean of n_2 and n_3.

Results

To determine the non DE genes for the construction of the baseline distribution of D we plotted (m, $Var(|\overline{D}|_{(m)})$) for m=1, ..., p, where $|\overline{D}|_{(m)}$ denoted the m-th order statistic of $\{|\overline{D}^{(g)}|\}_{g=1}^{p}$. From Fig. 4 we concluded that the first 100 order statistics provide information on the non-DE genes, which constitutes N_ε.

Figure 1. The plot of (m, $Var(|\overline{D}|_{(m)})$) for m=1, ..., 500.



We observed from Figure 1 that beyond the 100th smallest genes (in terms of $|\overline{D}|$), the variance tended to slowly increase. It is interesting to note that the first smallest 100 genes show rather

unstable variances. Based on these 100 genes we could estimate δ_0 and σ_0^2.

Once distributions of D and D_0 are determined, one can proceed calculating $ROC_{pair}(t_0)$ with a suitably chosen c_0 value which, in turn, determines the false positive rate t_0 in the baseline distribution. In general c_0 is chosen to make t_0 very small. However, as Pepe *et al.* (2003) indicates, with small number of tissue specimens, estimation of $ROC(t_0)$ at very small t_0 is not possible and hence in the real application one needs to compromise in the chose of t_0 such that it is small, but large enough to make $ROC(t_0)$ reasonably precise. Our choice of t_0 is 1/36. The top 20 genes in terms of $ROC_{pair}(1/36)$ values have 50 % overlap with top 20 genes in terms of t statistic and this result is exhibited in Table 2. The highlighted part in Table 2 indicates that the top gene in terms of the t statistic was ranked the 9-th in terms of $ROC_{pair}(1/36)$.

Table 3 shows $ROC_{mix}(1/36)$ values based on Eq. (3) and their corresponding ranks in terms of t_3 statistic of Kim *et al.* (2004).

Discussion

We have developed statistical methods which are applicable to the mixed data set of microarray experiment performed on human cancers by extending the ROC approach. The mixed data set of Table 1 occurs quite often in practice when the tissue material is not large enough to yield the adequate amount of RNA for undergoing the DNA microarray experiment. There is a possibility that the ROC approach, in general, and the ranking the genes, in particular, might be sensitive due to the small sample size. A study of sensitivity of this ROC approach by random sampling needs to be carried out and it is proposed as the future research.

Acknowledgements

BS Kim's study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ1-PG3-10411-00-03). SH Lee's work was supported by grant R04-203-000-10145-0 from the Basic Research Program of the Korea Science and Engineering Foundation. SY Rha's works were supported by the Korea Science and Engineering Foundation (KOSEF) through the Cancer Metastasis Research Center (CMRC) at Yonsei University College of Medicine

References

Kim BS, Kim I, Lee S, Rha SY, Hyun CH. (2004). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, in press.

Pepe MS. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction,* Oxford University Press, New York.

Pepe MS, Longton G, Anderson GL, Schummer M. (2003). Selecting differentially expressed genes from microarray experiments, *Biometrics,* **59**, 133-142.

Simon RM, Radmacher MD, Dobbin K. (2002). Design of studies using DNA microarrays. *Genetic Epid.* 23:21-36.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai

J, Speed TP. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, **30**(4), e15.

Table 2. Top 20 genes in terms of ROC are compared against their corresponding ranks in terms of t statistic. Highlighted part is the top gene in terms of t statistic, but it ranked the 9th in ROC

Rank by ROC_{pair}	Rank by t-stat	ROC_{pair} (1/36)	t-stat
1	4	1	-18.49
2	2	0.97	-20.89
3	8	0.97	17.13
4	12	0.97	16.91
5	17	0.97	-15.94
6	47	0.97	-13.81
7	80	0.97	-12.72
8	83	0.97	-12.64
9	1	0.94	22.30
10	6	0.94	-17.78
11	20	0.94	15.63
12	23	0.94	15.48
13	32	0.94	-14.40
14	66	0.94	13.07
15	81	0.94	-12.71
16	14	0.92	16.31
17	38	0.92	-14.24
18	42	0.92	-14.11
19	60	0.92	13.31
20	3	0.89	18.92

Table 3. Top 20 $ROC_{mix}(1/36)$ values and their corresponding ranks in terms of t_3 statistic of Kim *et al.* (2004).

Rank	$ROC_{mix}(1/36)$	Rank in terms of t_3
1	1.0000	2
2	1.0000	12
3	0.9833	1
4	0.9833	5
5	0.9833	25
6	0.9750	7
7	0.9750	10
8	0.9708	4
9	0.9708	23
10	0.9708	28
11	0.9708	40
12	0.9625	3
13	0.9625	98
14	0.9583	31
15	0.9542	8
16	0.9542	9
17	0.9542	50
18	0.9500	36
19	0.9500	77
20	0.9458	13